Minimax Theory and its Applications

Volume 11 (2026), No. 4s, 1065-1079

Saddle-Point Certified Training via Cone-Relaxed Duality: Neural Network Verification against Adversarial Perturbations in Medical Imaging

Nithya DeviS M[1,*], Naduvathezhath Nessariose Jose [2], Vivekanandan G[3]

[1]Assistant Professor, Department of Computational Intelligence, SRM institute of science and Technology, Chennai, India.
[2]Senior Consultant, Denken Solutions, Irvine, California, United States
[3]Assistant professor, Department of Computer Science Engineering, Sri Sairam Engineering College, Chennai, India.
Email: nithyads2@srmist.edu.in, josenaduvan@gmail.com, vivekanandan.nba.it@gmail.com

Abstract
Deep networks for chest X-ray diagnosis can change their predictions under small pixel perturbations, so formal robustness guarantees are essential for safe clinical use. Existing sparse second-order cone verifiers certify networks only after training, select neuron couplings with hand-made heuristics, and are benchmarked mainly on MNIST and CIFAR-10 with small test subsets. Deterministic certified robustness on lightweight biomedical images therefore remains largely unexplored, and this study aims to train and certify a medical image classifier within one saddle-point framework. Experiments use PneumoniaMNIST from MedMNIST+, which holds 5,856 chest X-ray images at 28×28 pixels with official training, validation, and test splits. The proposed Saddle-point Kolmogorov-Arnold Coupling Certified Network (SKAC-CertNet) casts training as a minimax problem in which a sparse second-order cone relaxation acts as a differentiable inner maximizer. A Kolmogorov-Arnold coupling layer learns which neuron pairs receive cone constraints through soft top-k gates under a fixed budget, and its dual-based margin bound enters the training loss directly. Weights and dual variables are updated by a primal-dual loop supported by a validity and monotone-tightening lemma and a convergence analysis, and the exact sparse cone program certifies each test image. Evaluation covers balanced accuracy, certified-robust accuracy at $\ell\infty$ radii from 1/255 to 8/255, certified radius, bound gap, and solver time against LP, SDP, and branch-and-bound verifiers, with the full model achieving 95.91% balanced accuracy (BAcc) and 86.0% certified-robust accuracy (CRA) at $\varepsilon = 2/255$. The study contributes a training-time certified framework that links saddle-point duality with learned cone selection, together with a deterministic certified-robustness benchmark on MedMNIST+.

Keywords: Certified robustness; Saddle-point optimization; Second-order cone relaxation; Neural network verification; Medical image classification

I. Introduction

Convolutional Neural Networks (CNNs), one type of deep learning architecture, are being used more and more in medical image processing for automated illness classification, diagnostic interpretation, and clinical decision-making(Ma et al., 2024). Although neural networks can classify images, they can be susceptible to adversarial perturbations, where small and imperceptible image changes can lead to incorrect predictions (Xiao et al., 2024). In biomedical context, these errors could result in misdiagnosis and affect therapeutic choices. For this reason, it has become a research field of particular interest to design mathematically robust systems of artificial intelligence to ensure the trustworthiness of these systems (Singh, Singh, et al., 2025). Several existing methodologies are under consideration such as adversarial training, interval bound propagation, linear programming relaxation, semidefinite programming, and formal neural network verification (Mozhegova et al., 2025). Convex relaxation is one of these methods that enables deterministic robustness certification beyond empirical attack testing by offering mathematical upper bounds on worst-case adversarial behavior. Recently, linear and semidefinite programming relaxations have been developed that further enhance the verifier's tightness and its computational scalability (Tobler et al., 2025). These developments lay the groundwork for the construction of safe and reliable neural network architectures that deliver consistent predictions in presence of bounded adversarial perturbations, especially in safety-critical biomedical image classification tasks.

There are several unresolved drawbacks in existing research, despite the improvement of the adversarial robustness verification(Dhar et al., 2026). Most existing approaches for verifying order two cones are based on the conventional sparse second-order cone verification methods, which certificate the network after training, and do not directly affect the optimization of weights. In addition, the choice of which interactions to select for tighter relaxation is often based on predetermined coupling structures or on computational heuristics, and thus, verification constraints are not very flexible(Z. Zhang et al., 2026). While certified training methods based on

intervals and relaxation can yield conservative bounds, semidefinite programming methods are also quite complex and are not well suited for modelling large inter-neuron dependencies. Recent certified training studies have also shown that tight convex relaxations can result in non-smooth and discontinuous training goals, which are challenging to train effectively(Sarih et al., 2025). The saddle-point formulation suggested in the paper is not proposed by Gaussian loss smoothing, however, it does help to solve some optimization problems(Taheri et al., 2026). Moreover, most medical imaging robustness studies conducted to date focus on the empirical defence against adversarial attacks, instead of a mathematical certification(Srinivasan et al., 2026). Thus, it is essential to develop an integrated framework of adaptive neuron-coupling selection, differentiable dual-based certification and theoretically supported primal-dual optimization as an important research line for lightweight biomedical image classification.

This study introduces a deterministic certified training framework called SKAC-CertNet (Saddle-point Kolmogorov-Arnold Coupling Certified Network) to construct an adversarially robust certified biomedical image classification model using a sparse second-order cone relaxation. The model is built on a light-weight convolutional backbone and a Kolmogorov-Arnold Coupling (KAC) layer, which represents the interaction between neurons. A sparse second-order cone program (SOCP) is constructed using the ReLU convex-hull constraints, selective McCormick envelopes, and second-order cone forms of the two-by-two positive-semidefinite minors for the inner maximization of the adversarial margin. Under fixed computational budget, learnable soft top-k gates determine which pairs of neurons are given those constraints. The relaxed problem is formulated in both primal and dual forms and the dual form provides a differentiable certification bound which in turn becomes part of the training objective. A primal-dual algorithm alternately updates the network weights and the dual variables, with the help of a validity analysis and convergence conditions. The goal is to increase the accuracy of certified robustness without compromising classification accuracy and computational speed.

A. Research Motivation

To achieve reliable robust performance in biomedical image classification under small perturbations, the classifiers must be robust to adversarial attacks. But the current methods for verification have loose convex relaxation bounds, are computationally expensive to optimize the SDP and don't have a tight connection between training the network and verifying robustness. Furthermore, the selection of neurons by heuristics limits the adaptive constraint minimization. Accordingly, this work introduces a new method, called SKAC-CertNet, which combines saddle-point optimization, sparse Second-Order Cone Programming (SOCP) relaxation and adaptive neuron-coupling selection, to obtain more compact deterministic robustness certificates without sacrificing the classification accuracy of biomedical images.Finally, this study proposes a new method, SKAC-CertNet, consisting of saddle-point optimization, sparse Second-Order Cone Programming (SOCP) relaxation, and adaptive neuron-coupling selection, to obtain tighter deterministic robustness certificates with an improved computational efficiency and reliable biomedical image classification.

B. Significance of the Study

This work proposes a mathematical framework to enhance the reliability of deep learning-based biomedical image classification in the presence of adversarial perturbations. The proposed SKAC-CertNet combines sparse SOCP relaxation, saddle-point optimization and adaptive neuron coupling to set up deterministic robustness certificates during training. Its primal-dual optimization method is to enhance the tightness of certification and to decrease the computational complexity. The work also helps identify systematical robustness evaluation, increase the mathematical transparency and improve the development of reliable and computational efficient deep learning systems for biomedical image analysis.

C. Problem Statement

Adversarial perturbations, which involves minor modifications to the input, can generate wrong classifications and affect the accuracy of the diagnosis in deep learning-based biomedical image classification systems. Current verification techniques for neural networks have mathematically sound guarantees of robustness, but are based on loose convex relaxations, are computationally prohibitive due to semidefinite optimization, and are separate from network training and post-training certification(Singh, Sarita, et al., 2025). Moreover, prior to neuron coupling selection, a predefined structure or heuristic has to be considered, which restricts adaptive constraint selection. Relaxation objectives that are not smooth also present optimization challenges to the existing certified training approaches(Pal et al., 2024). Therefore, a unified framework which optimizes network weights, worst-case adversarial perturbations and adaptive neuron-coupling constraints is still required. To overcome these challenges, this work suggests the use of saddle-point optimization, sparse Second-Order Cone Programming (SOCP) relaxation, learnable soft top-k coupling selection and differentiable dual-based certification to provide a deterministic adversarial robustness for light-weight biomedical image classification.

D. Key Contributions

- Deterministic robustness guarantees against perturbations protect pneumonia chest X-ray image classifiers, where small changes alter diagnoses.
- Training and certification share one minimax objective, so weights and certificates improve together throughout training epochs.
- SKAC-CertNet couples Kolmogorov-Arnold layers with a sparse second-order cone relaxation acting as the differentiable inner maximizer.

- Soft top-k gates pick couplings under a fixed budget and dual margin bounds drive primal-dual updates.
- PneumoniaMNIST results report 95.91% balanced accuracy (BAcc) and 86.0% certified-robust accuracy (CRA) at ε = 2/255, with supporting validity proofs.

E. Rest of the paper

The rest of the paper follows the organization: The related works on certified training and neural network verification are reviewed in section 2. The proposed SKAC-CertNet methodology is presented in Section 3. A description of the experimental setup is given in section 4. Results and discussion are in Section 5. Section 6 concludes the study and outlines future work.

II. Related Works

Zhang et al. (2024) proposed GaLileo, a general linear relaxation framework to improve robustness certification in Transformer architectures. The study covered the shortcomings of the traditional linear relaxation approaches in modeling the dependencies among the multidimensional nonlinear operations. The idea of an n-dimensional relaxation approach was used to efficiently represent complex nonlinear functions such as softmax operations. Experiments on SST and Yelp datasets showed certified radii up to 3.24 times greater than the baseline CROWN-BaF approach with comparable computational costs. The framework, though, was oriented toward the verification; not toward joint optimization of network parameters and robustness certificates. Shi et al. (2025) proposed a Branch and Bound algorithm for verifying neural networks with general nonlinear activation functions (GenBaB). This method involved linear bound propagation and an enhanced neuron-branching heuristic—both to lower the complexity of the verification process. Experimental testing showed that it can be used with sigmoid, tanh, sine, GeLU and multidimensional nonlinear operations. However, the approach mainly increased the efficiency of the verification process without incorporating differentiable certificates into the network training. Kabaha and Cohen (2024) explored formal verification of global robustness properties for neural network classifiers. They have developed a framework to expand the robustness analysis of individual input perturbation neighbourhoods to wider input neighbourhoods. The study helped to improve the formal robustness verification, but did not cover network training certified by the network or convex relaxation. Anderson et al. (2025) proposed a branching-based method that enhances the tightness of LCP and SDP relaxations. They followed such a method by dividing the uncertainty regions of the input and setting theory-based branching strategies and they evaluated on MNIST, CIFAR-10 and Wisconsin breast cancer diagnosis datasets. The framework, however, focused on verifying the post-training weights, and not on adaptive saddle-point optimization and network weight learning.

Chiu et al. (2025) proposed a hybrid verification framework, SDP-CROWN, that is a fusion of two strengths of SDP: tightness and scalability of linear bound propagation. The approach adds inter-neuron coupling information to linear bounds with just one extra value per layer. Experiments showed that verification is improved when the neural network used is large, with up to 65,000 neurons. The framework is however, mainly based on post-training verification, and it does not attempt to optimize the parameters of the network and the adaptive coupling constraints together. To evaluate the certified training algorithms under uniform settings, Mao et al. (2024) introduced CTBench, a benchmarking library. It tackles the inconsistency in training schedules, certifications process and selection of hyper parameters of existing methods. They found that such standardized evaluation has a profound effect on reported certified accuracy and that there are certain properties of certified models such as sparsely activated networks and regularization effects. However, CTBench does not introduce a new sparse SOCP-based training architecture but only offers benchmarking infrastructure. Balauca et al. (2024) proposed Gaussian Loss Smoothing (GLS) to overcome problems of optimization with tight convex relaxations. They presented a zeroth-order optimization algorithm, Policy Gradients with Parameter-based Exploration (PGPE), and a first-order optimization algorithm, Randomized Gradient Smoothing (RGS). It was shown that smoothing of objectives improved certified robustness in experiments. The framework, however, does not explore sparse SOCP-based saddle-point optimization and adaptive neuron-coupling selection. Maleki et al. (2026) proposed a sparse second-order cone programming approach to efficiently certify the robustness of neural networks. They use a combination of the 2×2 principal minors derived second order cone constraints, the ReLU convex-hull relaxation and selective McCormick lifting. It increases the tightness of the certificate but it solves the problem of post-training verification, not of learning the weights of the networks and the couplings of the adaptive neurons together with them in the course of training by means of the saddle-point optimization. Sparse second-order cone verifiers are only certified for networks after their training, rely on handcrafted coupling heuristics, and have been evaluated primarily on MNIST and CIFAR-10, but have not been evaluated on medical images deterministically. The solution is called SKAC-CertNet, which trains weights and certificates simultaneously, learns cone couplings using soft gates and deterministically certifies chest X-rays.

III. Saddle-Point Certified Training Methodology for Pneumonia Image Classifiers

This section introduces us to a saddle-point certified network for pneumonia detection in chest X-ray images: SKAC-CertNet. There are three interrelated components to the method. In part one, the images of the PneumoniaMNIST dataset and the perturbation boxes are created. In part two, a compact ReLU network,

consisting of a Kolmogorov–Arnold Coupling (KAC) layer is built. In part three, the network is trained as a minimax game, and certified using a sparse second order cone (SOC) relaxation. Fig. 1 shows the pipeline. The results produced in each step are used as inputs to the next step. Robustness is a minimax perspective. The learner makes a minimum of a loss. The opponent makes use of it in a small (pixel) box. ReLU networks are not the easiest to train, in particular not on the inner maximum, thus the convex cone relaxation provides a safe upper bound. This is a training signal and final certificate. The same amount of trains and verifies the network.

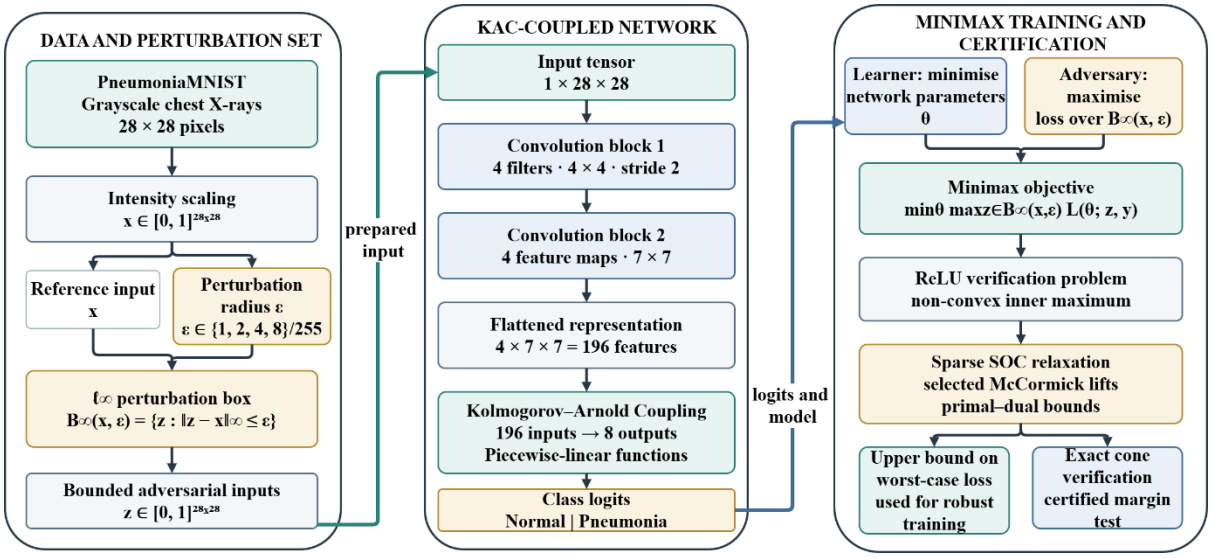

Fig. 1. SKAC-CertNet framework: (I) data scaling and ℓ∞ box construction, (II) compact ReLU backbone with Kolmogorov–Arnold Coupling layer, (III) minimax training with sparse SOC relaxation and exact cone certification.

A. Data and Preprocessing

PneumoniaMNIST is a subset of MedMNIST v2(Jiancheng Yang et al., 2024). It holds 5,856 chest X-ray images of size 28 × 28. Each image is labelled normal or pneumonia. The official split gives 4,708 training, 524 validation, and 624 test images. The data are public, carry a CC BY 4.0 licence, and are meant for method research, not clinical use. Only the 28 × 28 size is used, so that cone problems stay solvable. Small inputs also keep the number of ReLU units low. Raw grey levels I_i are scaled and clipped as defined in Eqn.(1)

$$x_i = \Pi\left(\frac{I_i}{255}\right), \qquad \Pi(z) = \min\{\max\{z, 0\},1\} \tag{1}$$

Where,$i = 1, \ldots, 784$ and Π is the clipping operator. The same range defines the perturbation box later. The scaled image is the vector x. Pneumonia images usually outnumber normal ones. Class weights offset this imbalance, is calculated using Eqn. (2):

$$\pi_c = \frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} \mathbb{1}\,[y_i = c], \qquad w_c = \frac{1}{2\pi_c} \tag{2}$$

Here,π_c is the share of class c, and $\mathbb{1}[\cdot]$ is the indicator function. Cone solving is slow, so a fixed random subset J of $M = 200$ test images is drawn once with a stored seed. Every subset has the same chance,as defined in Eqn. (3):

$$\Pr(J) = \frac{M!\,(N_{\text{test}}-M)!}{N_{\text{test}}!} \tag{3}$$

The threat model is the ℓ_∞ ball of radius ε around a clean image x_0. Every pixel stays inside [0,1],as defined in Eqn. (4):

$$\Delta_\varepsilon(x_0) = \{\, x:\ \| x - x_0 \|_\infty \le \varepsilon,\ 0 \le x \le 1 \,\} \tag{4}$$

Here,ε is the attack limit. The double bars give the largest absolute pixel change. The pixel limits of this set, as defined inEqn.(5)

$$\underline{x_i} = \max\{x_{0,i} - \varepsilon,\ 0\}, \qquad \overline{x_i} = \min\{x_{0,i} + \varepsilon,\ 1\} \tag{5}$$

The test sweep uses $\varepsilon \in \{1,2,4,8\}/255$. Training uses $\varepsilon = 2/255$.

B. Compact Backbone and KAC Layer

The backbone has two convolution blocks. Each block has 4 filters of size 4 × 4, stride 2, and padding 1. The first block gives 4 maps of size 14 × 14. The second gives 4 maps of size 7 × 7, which is 196 values. Max pooling is avoided because it loosens the bounds. Each block is an affine map followed by a ReLU Eqn.(6):

$$s^{(k)} = W^{(k)} z^{(k-1)} + b^{(k)}, \qquad z^{(k)} = \max\{s^{(k)}, 0\}, \qquad z^{(0)} = x \tag{6}$$

Here,$s^{(k)}$ is the pre-activation, $z^{(k)}$ is the activation, and $b^{(k)}$ is the bias. $W^{(k)}$ is the sparse matrix form of the filters. It repeats the same filter weights. Interval bounds come from the box in Eqn. (5). Let $W_+ = \max\{W, 0\}$, $W_- = \min\{W, 0\}$, and $[\cdot]_+ = \max\{\cdot, 0\}$. Then in Eqn.(7)

$$\begin{aligned} l^{(k)} &= W_+^{(k)}\left[l^{(k-1)}\right]_+ + W_-^{(k)}\left[u^{(k-1)}\right]_+ + b^{(k)}, \\ u^{(k)} &= W_+^{(k)}\left[u^{(k-1)}\right]_+ + W_-^{(k)}\left[l^{(k-1)}\right]_+ + b^{(k)} \end{aligned} \tag{7}$$

Where, $\left[l^{(0)}\right]_+ = \underline{x}$ and $\left[u^{(0)}\right]_+ = \overline{x}$. These bounds start the relaxation. They also decide which ReLUs are stable. Stable ReLUs need no extra constraint.The KAC layer replaces fixed weights with learnable piecewise-linear edge functions. Inputs are clipped to $[0,1.5]$. This range has 3 intervals of width $\mu = 0.5$ and knots $t_m = m\mu$. The edge function from input i to output j, as defined in Eqn.(8)

$$\varphi_{ji}(t) = \sum_{m=0}^{3} \zeta_{jim}\ \Lambda_m(t), \qquad \Lambda_m(t) = \max\left\{0,\ 1 - \frac{|t-t_m|}{\mu}\right\} \tag{8}$$

The layer has 196×8 edge functions, so it holds 6,272 learnable heights ζ_{jim}. Each edge function is rewritten with two ReLU units at the inner knots. Hence the layer stays a ReLU network, and the cone hulls apply without change. A usual layer multiplies each input by one fixed weight. The KAC layer lets each connection learn its own shape. The 8 outputs and the class scores are calculated using Eqn.(9):

$$\begin{aligned} h_j &= \max\{0,\ \textstyle\sum_{i=1}^{196} \varphi_{ji}\left(\min\{z_i, 1.5\}\right) + \beta_j\}, \\ f(x) &= W_o\, h(x) + b_o, \qquad \hat{y} = \operatorname*{argmax}_c f_c(x) \end{aligned} \tag{9}$$

Cone constraints are costly. A budget of $B = 32$ pairs per layer is split into 16 cross-layer, 8 current-layer, and 8 previous-layer pairs. Each candidate pair (p, r) has a learnable score a_{pr} and a soft gate, as shown in Eqn. (10)

$$g_{pr} = \sigma\left(\frac{a_{pr} - \tau_B}{\omega}\right), \qquad \sigma(u) = \frac{1}{1+e^{-u}} \tag{10}$$

Where, τ_B is the B-th largest score and $\omega = 0.1$. A score above τ_B gives a gate near 1. A lower score gives a gate near 0. The small temperature keeps the gate close to a hard choice while gradients still flow. At test time, the 32 highest scores are kept. The whole network has about 6,600 trainable values.

C. Minimax Training Formulation

Training is a zero-sum game. The learner picks θ, which holds filters, edge heights, biases, and gate scores. The adversary picks the worst image in $\Delta_\varepsilon(x_0)$. With class weights w_y and cross-entropy ℓ, the game is defined in Eqn.(11).

$$\min_\theta \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} w_{y_i} \max_{x \in \Delta_\varepsilon(x_{0,i})} \ell(f_\theta(x), y_i) \tag{11}$$

With two classes, only one rival class $1 - y$ exists. The worst-case margin is defined in Eqn.(12)

$$m_\theta^*(x_0) = \max_{x \in \Delta_\varepsilon(x_0)} \left[f_\theta(x)_{1-y} - f_\theta(x)_y\right] \tag{12}$$

A negative margin means that no allowed change flips the prediction. Hence the sign of m^* is the robustness test. The inner problem is non-convex, so strong minimax equality cannot be assumed. Only the weak minimax inequality always hold,as expressed in Eqn.(13)

$$\sup_x \inf_\theta\ \ell(\theta, x)\ \le\ \inf_\theta \sup_x\ \ell(\theta, x) \tag{13}$$

Therefore, m^* is replaced by a computable upper bound m^u. The loss mixes clean and certified terms, as defined in Eqn.(14)

$$L(\theta) = \frac{1}{N_{\text{train}}} \sum_i w_{y_i} \left[(1-\kappa)\ \ell\left(f_\theta\left(x_{0,i}\right), y_i\right) + \kappa \log\left(1 + e^{m_i^u}\right)\right] \tag{14}$$

The softplus term is small for a negative bound and grows almost linearly for a positive one. The mixing weight κ lies between 0 and 1. A warm-up keeps early training stable, because a loose bound at the start would dominate the loss. The weight κ and the radius ε rise linearly over the first 10 of 30 epochs, as specified in Eqn.(15)

$$\kappa_e = 0.5 \cdot \min\{1,\ e/10\}, \qquad \varepsilon_e = \frac{2}{255} \cdot \min\{1,\ e/10\} \tag{15}$$

D. Sparse Second-Order Cone Relaxation

A ReLU with $l_j \ge 0$ or $u_j \le 0$ is stable and is treated exactly. An unstable ReLU has $l_j < 0 < u_j$. It is replaced by its triangle hull, as formulated in Eqn.(16)

$$\max\{0, s_j\}\ \le\ z_j\ \le\ \frac{u_j\left(s_j - l_j\right)}{u_j - l_j} \tag{16}$$

The hull treats each neuron alone, so it ignores links between neurons. To restore some links, selected products $z_p z_r$ are lifted to variables η_{pr}. McCormick planes bound each product, using $0 \le z_p \le u_p$ and $0 \le z_r \le u_r$,as expressed in Eqn.(17)

$$\max\{0,\ u_r z_p + u_p z_r - u_p u_r\}\ \le\ \eta_{pr}\ \le\ \min\{u_r z_p,\ u_p z_r\} \tag{17}$$

Let ξ_p stand for z_p^2. It is tied to z_p bylifted matrix relationship defined in Eqn.(18)

$$z_p^2\ \le\ \xi_p\ \le\ u_p z_p \tag{18}$$

The 2×2 minor of the lifted matrix is a second-order coneas represented in Eqn.(19)

$$\eta_{pr}^2 \le \xi_p \xi_r \quad \Leftrightarrow \quad \left\|\left(2\eta_{pr},\ \xi_p - \xi_r\right)\right\|_2 \le \xi_p + \xi_r \tag{19}$$

Only gated pairs receive Eqn. (17)–(19). The vector q has $+1$ at the rival class and -1 at the true class. The relaxed problem is formulated in Eqn.(20)

$$m_{\text{SOCP}}^u(x_0) = \max_{v \in \mathcal{F}(x_0;\, \mathcal{E}, \mathcal{C}, \mathcal{P})} q^\top (W_o\, h + b_o) \tag{20}$$

Where,v holds all variables and $\mathcal{F}$ is the feasible set. The value m^u_{SOCP} is never below the true margin. Each added pair can only keep it equal or lower it. Only a few pairs are lifted, so the problem is far smaller than a full semidefinite lift. The pair budget controls the trade-off between bound tightness and solve time. A larger B gives a tighter bound but a slower solve. The pair sets $\mathcal{E}$, $\mathcal{C}$, and $\mathcal{P}$ are the cross-layer, current-layer, and previous-layer pairs.

E. Dual Surrogate and Primal–Dual Updates

A cone solve at every step is too slow. Weak duality gives a faster path. For any $\lambda \geq 0$,as established in Eqn. (21)

$$m^*_\theta(x_0) \;\leq\; m^u_{\mathrm{SOCP}}(x_0) \;\leq\; D(\lambda;\theta) \tag{21}$$

A feasible λ comes from a backward pass. Here $\Gamma^{(k)}$ is the diagonal matrix of ReLU slopes. An active neuron has slope 1, an inactive neuron has 0, and an unstable neuron has $u_j/\big(u_j - l_j\big)$defined in Eqn. (22)

$$\lambda^{(k)} = \Gamma^{(k)}\big(W^{(k+1)}\big)^{\top}\lambda^{(k+1)}, \qquad \lambda^{(K+1)} = q \tag{22}$$

The differentiable bound used inside the loss is formulated in Eqn. (23)

$$\begin{aligned} D(\lambda;\theta) \quad &= \nu^{\top}x_0 + \varepsilon \parallel \nu \parallel_1 + d - \textstyle\sum_{(p,r)} g_{pr}\ \rho_{pr}, \\ \nu \quad &= \big(W^{(1)}\big)^{\top}\lambda^{(1)} \end{aligned} \tag{23}$$

Where,d collects bias and hull terms, and $\rho_{pr} \geq 0$ is the gain of pair (p, r). This value is m^u in Eqn. (14). It has four parts: a clean-image term, a box term, a constant, and a gain for every selected pair. All parts are sums and products of trainable values, so the usual backward pass gives its gradient. Its cost is about two network passes per image, which suits batches of 64.The Lagrangian of the relaxed problem isdefined in Eqn. (24)

$$\mathcal{L}(v, \lambda; \theta) = q^{\top}(W_o\, h + b_o) - \lambda^{\top}\psi(v;\theta) \tag{24}$$

where $\psi \leq 0$ stacks all hull, plane, and cone constraints. The problem is convex. When a strictly feasible point exists, the saddle-point equality holds,as presented in Eqn. (25)

$$\max_{v}\min_{\lambda\geq 0}\ \mathcal{L}(v,\lambda;\theta) \;=\; \min_{\lambda\geq 0}\max_{v}\ \mathcal{L}(v,\lambda;\theta) \tag{25}$$

The loop starts from the closed-form λ of Eqn. (22). In each of $N_{\mathrm{in}} = 5$ inner steps, v climbs and λ descends, as formulated in Eqn. (26).

$$\begin{aligned} v_{n+1} \quad &= \Pi_V\big(v_n + \alpha_v\, \nabla_v \mathcal{L}(v_n, \lambda_n; \theta_t)\big), \\ \lambda_{n+1} \quad &= \max\{0,\ \lambda_n + \alpha_\lambda\, \psi(v_{n+1}; \theta_t)\} \end{aligned} \tag{26}$$

Here,Π_V projects the input part of v onto the box of Eqn. (5), and $\alpha_v = \alpha_\lambda = 0.05$. One weight step then follows, as expressed in Eqn. (27)

$$\theta_{t+1} = \theta_t - \alpha_\theta\, \nabla_\theta L(\theta_t, \lambda^{N_{\mathrm{in}}}) \tag{27}$$

Where,$\alpha_\theta = 10^{-3}$ and Adam rescales the step. Training uses batches of 64 images. The epoch with the best validation certified margin is kept.

F. Validity, Tightening, and Convergence

Three properties support the method. First, the bound is valid for every θ. Each exact point lifts to the relaxed set with $\xi = z^2$ and $\eta = z_p z_r$, as established in Eqn. (28)

$$m^*_\theta(x_0) \;\leq\; m^u_{\mathcal{E},\mathcal{C},\mathcal{P}}(x_0;\theta), \qquad \forall\, \theta \tag{28}$$

Second, adding pairs never loosens the bound when variable bounds stay fixedas demonstrated in Eqn. (29)

$$\begin{aligned} &\mathcal{E}_1 \subseteq \mathcal{E}_2, \quad \mathcal{C}_1 \subseteq \mathcal{C}_2, \quad \mathcal{P}_1 \subseteq \mathcal{P}_2 \\ &\Rightarrow\ m^u_{\mathcal{E}_2,\mathcal{C}_2,\mathcal{P}_2} \leq m^u_{\mathcal{E}_1,\mathcal{C}_1,\mathcal{P}_1} \end{aligned} \tag{29}$$

An added pair may give no gain, so the claim is not strict. Third, assume that L is smooth with constant L_s. Assume also that the inner step has a gradient error of at most δ. With $\alpha_\theta \leq 1/L_s$,characterized in Eqn. (30).

$$\min_{0\leq t<T} \parallel \nabla L(\theta_t) \parallel^2 \leq \frac{2(L(\theta_0) - L_{\min})}{\alpha_\theta\, T} + \delta^2 \tag{30}$$

ReLU networks are only piecewise smooth. Hence Eqn. (30) is a guide under stated assumptions, not a proof for every run. The first term shrinks as T grows. The second term is an error floor caused by the inexact inner step.

G. Exact Cone Certification

After training, the exact cone problem is solved for each image in J and each ε. It is built in CVXPY and solved by Clarabel on a CPU. The 32 top-scored pairs per layer are used. An image is certified when the prediction is correct and $m^u < 0$. The certified-robust accuracy iscalculated using Eqn. (31)

$$\mathrm{CRA}(\varepsilon) = \frac{1}{M}\textstyle\sum_{i\in J} \mathbb{1}\,[\hat{y}_i = y_i] \cdot \mathbb{1}[m^u_i(\varepsilon) < 0] \tag{31}$$

The certified radius of one image ismathematically defined in Eqn. (32)

$$\varepsilon^*(x_0) = \sup\{\varepsilon \geq 0:\ m^u_{\mathrm{SOCP}}(x_0;\varepsilon) < 0\} \tag{32}$$

A bisection search finds it to a resolution of $0.25/255$. The mean radius and the surrogate gap arecalculated using Eqn. (33).

$$\begin{aligned} \bar{\varepsilon} \quad &= \frac{1}{M}\textstyle\sum_{i\in J} \varepsilon^*(x_i), \\ \mathrm{Gap}(\varepsilon) \quad &= \frac{1}{M}\textstyle\sum_{i\in J}\big[D(\lambda_i;\theta) - m^u_{\mathrm{SOCP},i}(\varepsilon)\big] \end{aligned} \tag{33}$$

The margin reflects the accuracy of the fast surrogate. The relaxation is partial. Therefore, an uncertified image does not indicate that an adversarial attack took place. The solve time and memory for each image are recorded. There are three possibilities in the certificate. The image is certified if the margin is negative. It is misclassified

if the clean classification was incorrect. It is uncertified if the classification was correct but the margin was not negative. The separation allows the outcome to distinguish between errors in the model and constraints of verification. The scan over ε reveals the effect on the certified fraction.The unique aspect of SKAC-CertNet is the combination of training and verification in a single saddle-point framework. First, the second-order cone relaxation is sparse and is used as the differentiable inner maximizer, thus weights and certificates can be improved jointly. Second, the Kolmogorov-Arnold coupling layer, with 196 inputs and 8 outputs, is trained to determine the pairs of neurons that have a cone constraint using soft top-k gates within a budget of B = 32 pairs of neurons instead of handcrafted heuristic ways. Third, it is a dual-based margin bound whose proof is supported by a validity and monotone-tightening lemma and a convergence analysis. Finally, the following image of an exact sparse cone certification provides a deterministic robustness measure for chest X-rays in pneumonia.

IV. Experimental Setup

This section outlines the dataset, data preprocessing methodology, architecture and training methodology for SKAC-CertNet, and the verification methodology employed to test SKAC-CertNet. Chest X-ray images are classified as Normal or Pneumonia, using PneumoniaMNIST, which is in the collection of MedMNIST. The 28 x 28 version of the image is chosen to be compatible with the compact network layout. Official Training, Validation and Test Partitions are kept for Model Development and final evaluation. Images are normalized to [0, 1] and the adversarial robustness is explored as per ℓ∞ perturbation model. Each of the classification, certification, and computational metrics is evaluated separately with fixed evaluation settings and a fixed random seed. The summary of data set characteristics, preprocessing, model settings, perturbation bounds, certification protocol and evaluation measures are summarized in Table 1.

Table 1. Experimental configuration and evaluation settings for SKAC-CertNet framework.

Component	Experimental configuration
Dataset	PneumoniaMNIST
Classification task	Binary classification
Diagnostic classes	Normal and Pneumonia
Total images	5,856
Training set	4,708 images
Validation set	524 images
Test set	624 images
Image resolution	28 × 28 pixels
Input channels	1 (grayscale)
Input normalization	Pixel values scaled to [0, 1]
Network architecture	Compact convolutional backbone with KAC layer
Activation function	ReLU
KAC configuration	196 inputs × 8 outputs
Trainable parameters	Approximately 6,600
Threat model	ℓ∞-norm
Perturbation budgets	ε = 1, 2, 4, 8/255
Certification subset	200 fixed test images
Pair budget	B = 32
Verification formulation	Sparse SOCP relaxation
Optimization solver	CVXPY with Clarabel

V. Results and Discussion

The data representation, feature extraction, KAC coupling structure, minimax optimization, and formal verification pathway of SKAC-CertNet are discussed in this section. The discussion is about the proposed methodology and does not repeat the metric based plots. 8 method-related figures represent the internal processing stages and the mathematical relationship. All the classification, certification and computational performance measures are summarized in one comparative table. The last section deals with the functions of all the components of the individual frameworks and their interactions in the context of certification.

A. Input Representation and Perturbation Construction

PneumoniaMNIST is a dataset of grayscale chest X-ray images labled for classification as either pneumonia or no pneumonia. The input classes of SKAC-CertNet are Normal and Pneumonia types. The 784-dimensional vectors are the representations of each image, which were converted from their original two-dimensional arrangement. Pixel normalization sets up a range of numbers for the pixels and defines a valid input domain for subsequent optimization. The construction of the perturbation is done around each normalized image with an (L_∞)- bounded neighborhood. Lower and upper pixel limits are within the valid image range. This approach enables the adversarial optimization problem to work with bound changes to individual pixels while maintaining the validity of the input value. The compact network and the verification formulation are then informed of the

perturbation region. Thus, there is a common ground for minimax training and formal certification of the input representation and construction of threat regions.This means that there is a common basis for minimax training and formal certification of the input representation and construction of threat regions.

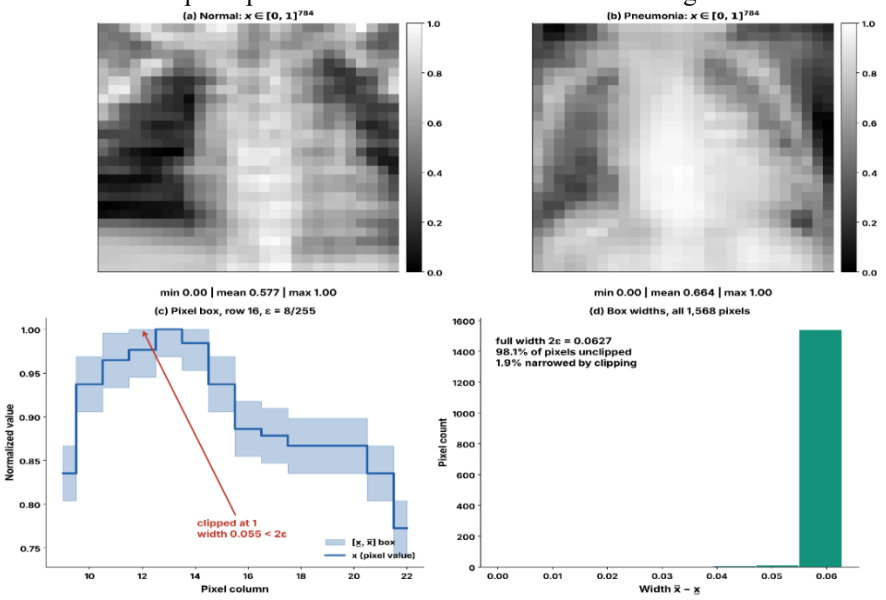

Fig. 2. Normalized chest X-ray inputs with clipped perturbation boxes.

Fig. 2 illustrates Normal and Pneumonia 28 × 28 image samples in panels (a) and (b), respectively, scaled using Eqn. (1), with mean values of 0.577 and 0.664 and a common intensity range of 0.00–1.00. In Panel (c), the width of the clipping box is reduced to 0.055 instead of 2ε and the plot shows one row of the box with ε = 8/255. The 1,568 pixel boxes are shown in panel (d), where the majority (98.1%) possess the full width of 0.0627, and 1.9% are narrowed, which constitutes the input region to be used for training and certification. The recovered 8-bit pixel intensities are displayed and all values are directly derived from this.

B. Hierarchical Feature Extraction

The compact convolutional backbone learns spatial representations from the normalized chest X-ray input. The first convolutional network layer detects local intensity variation and basic spatial features. The feature information to be used for binary classification is retained and the spatial dimensions are reduced in subsequent processing. The second convolutional block generates a compact representation of features. The output feature maps become a flattened 196 dimensional vector, which is fed into the KAC layer. This setup helps to decrease the dimensionality of the representation and the size of the following optimization problem. The compact feature pathway is relevant to the proposed certification strategy since the number of activation variables directly affects the complexity of the relaxation. The reduced representation is then used to link the architecture design to the computational needs of formal verification.

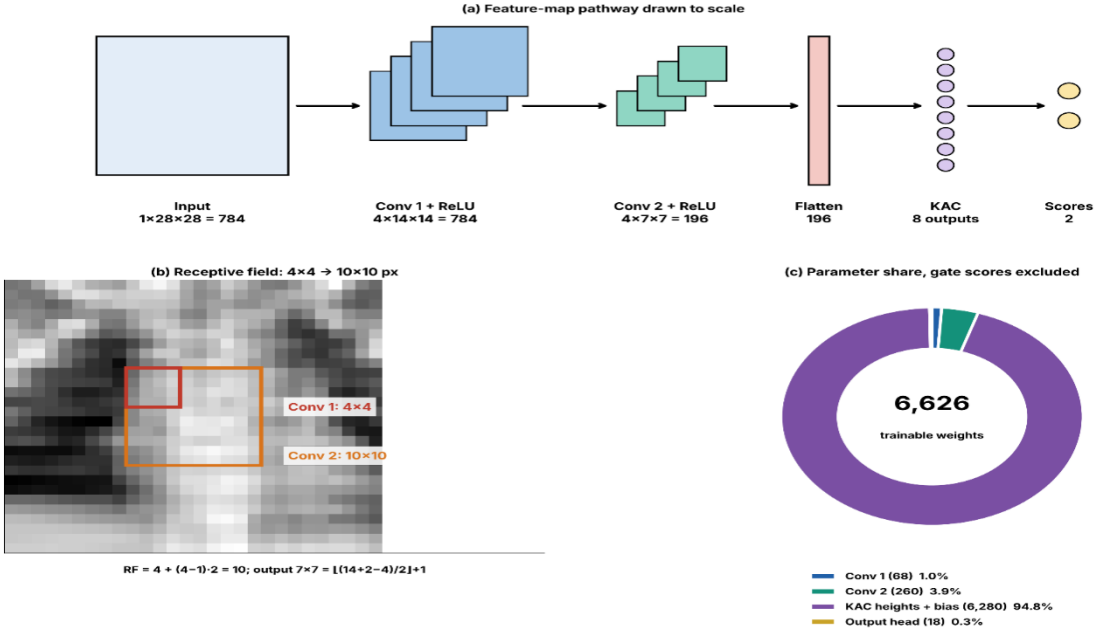

Fig. 3. Compact convolutional pathway, receptive field, and parameter share.

Fig. 3(a) shows that the tensors of dimensions 784 inputs, 4 × 14 × 14 feature maps, and 4 × 7 × 7 flattened to 196 units and 8 KAC outputs, while the final 2 class scores remain flat. In (b), the receptive field is overlaid, increasing from 4×4 to 10×10 input pixels. The KAC layer contains 6626 trainable weights, about 94.8% of which are in the KAC layer and the remaining 5% in the convolutions and output head, as shown in panel (c), indicating a compact, small relaxation. These panels illustrate why verification remains tractable: a smaller number of activations corresponds to a smaller number of lifted variables.

C. Kolmogorov–Arnold Coupling Representation

The compact feature representation is combined with the structured nonlinear interaction in the KAC layer. The coupling structure is different from a standard fully connected mapping, in that it is a representation of selected relationships between the input and output variables, mediated by parameterized univariate transformations. The flattened feature vector is then passed to the layer which produces the nonlinear representation needed for classification. It has a coupling structure that explicitly specifies selected input–output interactions. This representation is especially important for the verification formulation where the chosen interactions can be added to the sparse relaxation without having to build an unnecessarily dense interaction set. The KAC layer is thus linking the representational power of the network with the sparsity requirements of the minimax certification process.

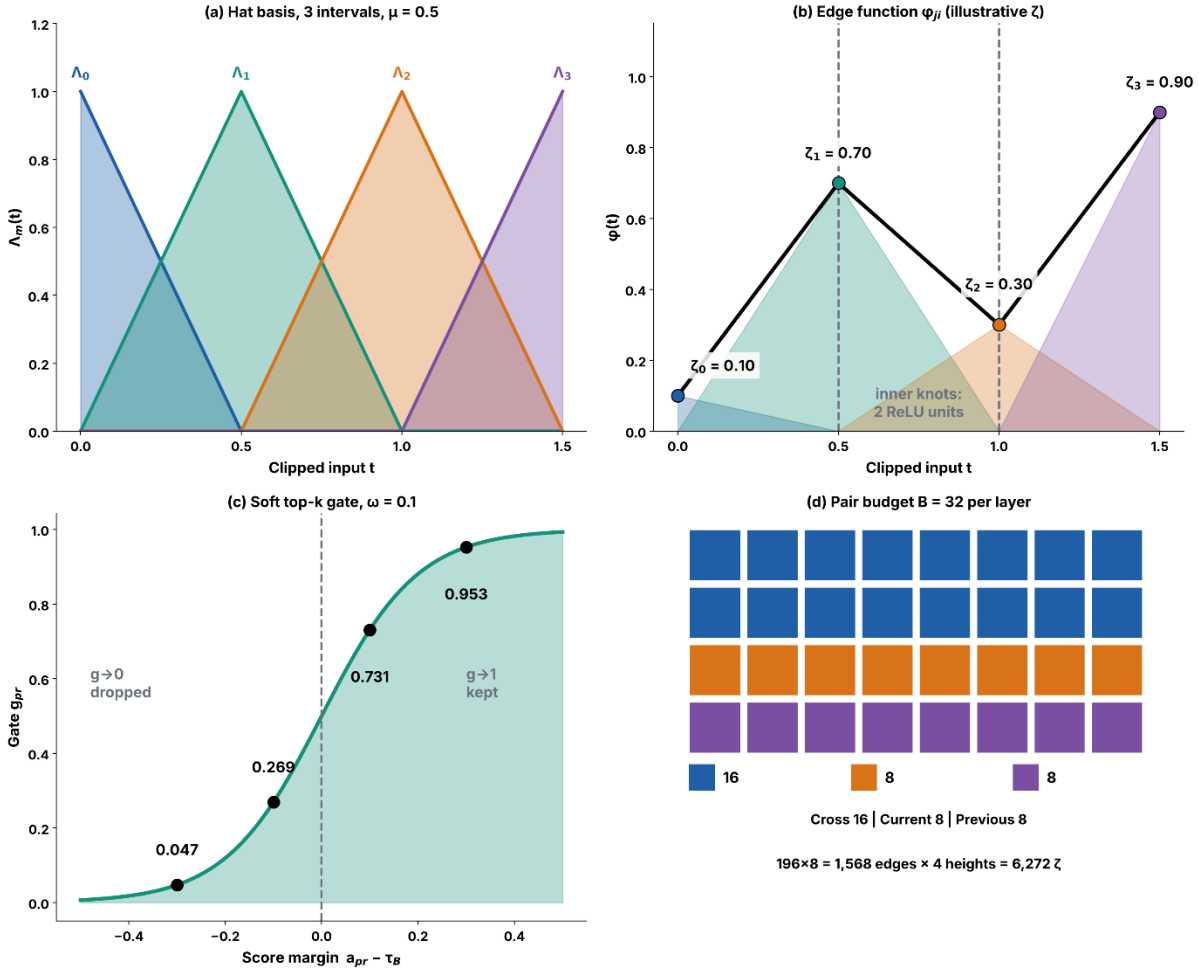

Fig. 4. KAC hat bases, soft gating, and pair budget.

Fig. 4 illustrates the four hat functions that are evaluated at four selected knots, namely 0, 0.5, 1.0 and 1.5, with μ = 0.5, in panel (a), and the combination of these four into a single piecewise-linear edge function in panel (b), with the inner knots being realized by two ReLU functions. Panel (c) shows the soft gate rising from 0.047 at −0.3 to 0.953 at +0.3 around the threshold. The number of cross-layer pairs in the panel is 16, the number of current-layer pairs in the panel is 8, and the number of previous-layer pairs in the panel is 8. The 32-pair budget is allocated in this manner.The 32-pair budget is divided into 16 cross-layer pairs, 8 current-layer pairs and 8 previous-layer pairs.

D. Minimax Optimization and Saddle-Point Structure

SKAC-CertNet reduces and solves the learning problem as a minimax optimization problem for network parameters and bounded adversarial perturbation. The learner tries to find network parameters for which the classification target is minimized while the adversarial variable tries to find perturbations for which the classification target is maximized within the limits of the input neighborhood. The robust training and the robust

verification are mathematically linked by the saddle-point formulation. The learner's and adversarial goals are expressed in the form of conflicting optimization directions. The assumptions of the optimization formulation impose the necessary optimality conditions for a valid saddle-point solution to be true. A relationship between the primal and dual optimization problems is given by the weak minimax inequality. Further regularity conditions are needed to get saddle-point equality, including strict feasibility when applicable.

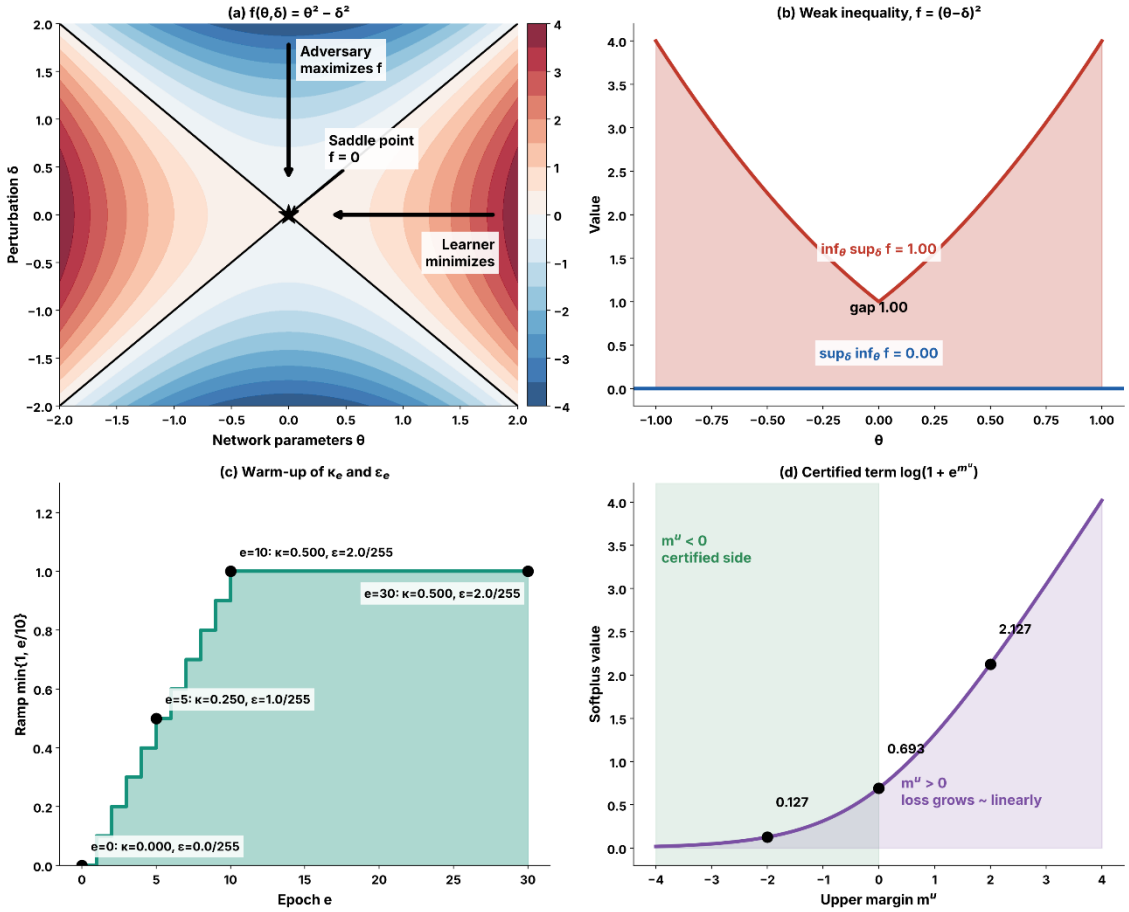

Fig. 5. Saddle geometry, weak inequality, warm-up schedule, and softplus.

Fig.5shows the saddle-point surface $(f = \theta^2 - \delta^2)$ at the origin, where the learner is descending and the adversary is ascending the surface. The weak inequality is displayed in panel (b) where sup-inf = 0.00 and inf-sup = 1.00. The ten-epoch warm-up, as seen in Panel (c), sets κ to 0.500, and sets the parameter ε to the value 2/255. In panel (d), the softplus certified term increased from 0.127 for a margin of −2 to 2.127 for a margin of +2, and got even stronger in that direction as the margin grew. Both schedules are ramped identically, and one curve is used to represent both quantities.

E. Sparse Second-Order Cone Relaxation

The problem of maximizing against an adversary in ReLU networks is hard to solve exactly due to these piecewise linear and coupled activation variables. A sparse second order cone relaxation is used to overcome this problem by SKAC-CertNet. Selected lifted variables and convex constraints are introduced in the relaxation to model relationships between the activations in the network. McCormick-type envelopes give linear outer approximations for some bilinear relationships, and second-order cone constraints keep more geometric structure. The pair control the number of selected interactions that are included in the relaxation. The feasible region is an easily solvable approximation of the original adversarial problem. The upper bound derived from this formulation is used to help with the certification process.

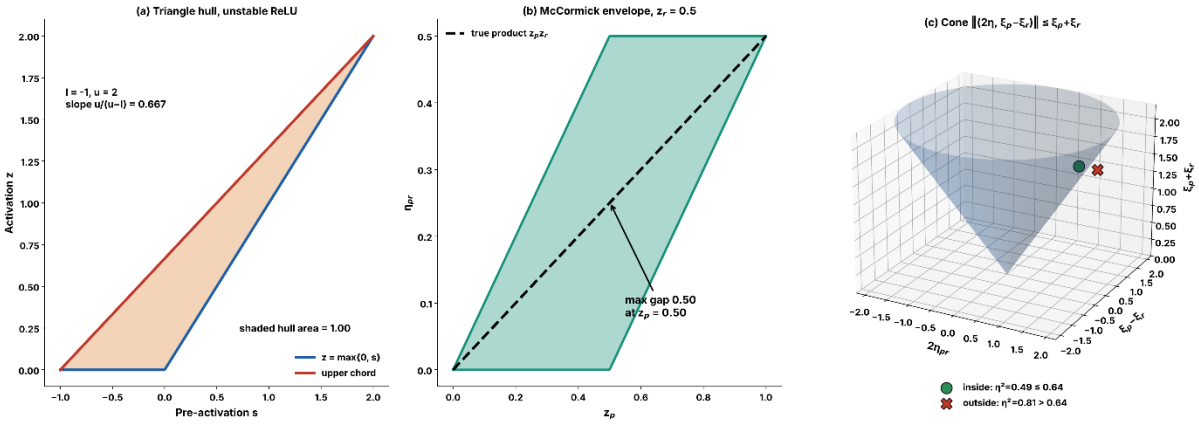

Fig. 6. Triangle hull, McCormick envelope, and second-order cone geometry.

Fig. 6(a) shows the triangular convex relaxation of an unstable ReLU function for l = −1 and u = 2, with the chord slope set at 0.667 and the relaxation area set at 1.00. In panel (b), the McCormick envelope for $z_r = 0.5$ is plotted, and is shown to have a maximum gap of 0.50 at $z_p = 0.50$. Panel (c) displays the second order cone, with a feasible lifted point (η^2 = 0.49) within the bound (0.64) of the cone, but η^2 = 0.81 is outside the cone. Both construction makes the entire problem convex.

F. Dual-Bound Certification and Verification Pathway

The dual formulation is a complementary observation of the relaxed adversarial optimization problem. By weak duality, if the dual objective $D(\lambda;\theta)$ is feasible, it is at or above the upper bound provided by the sparse SOCP relaxation, which means $m* \leq m_S^u OCP \leq D(\lambda;\theta)$. The difference between these values allows for the assessment of tightness of binding. The verification process is performed for each test image, and the output margin corresponding to each test image is compared within the specified perturbation region. Once the verified upper margin meets the negative-margin requirement of the competing class, it earns a certificate. Images failing that condition are not certified, but that does not always mean that the clean prediction is incorrect. The procedure for verification consists of the following steps: sparse SOCP problem construction, solver run, bound extraction, and checking the certification condition. If the solver fails, it is not counted towards a successful certification.Failure of the solver and failure to certify are considered distinct successes. This separation is crucial to ensure proper reporting of the success of the verification process.

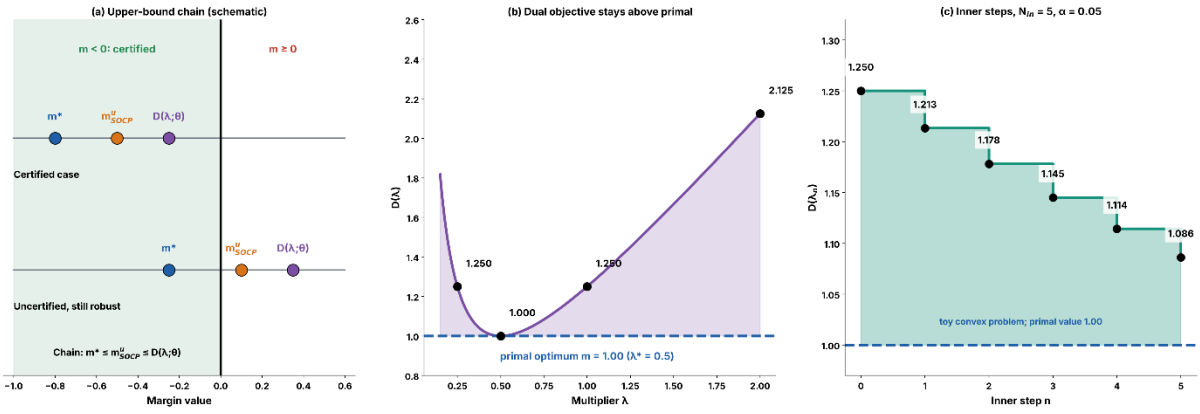

Fig. 7. Upper-bound chain, dual objective, and inner primal-dual iterations.

Fig. 7(a) shows ordering of the bounds as $m* \leq m_S^u OCP \leq D(\lambda;\theta)$. A certificate needs to have a negative upper margin; thus, a strong image could be rejected because the bound was not tight enough. Panel (b) is a toy problem whose dual objective is never less than the primal value 1.00, and attains it value at λ = 0.5. Five inner primal–dual steps (shown in Panel (c)) are reducing the value of the dual objective from 1.250 to 1.086, approaching the primal optimum. The toy problem is only an illustration and is not computed from the trained network.

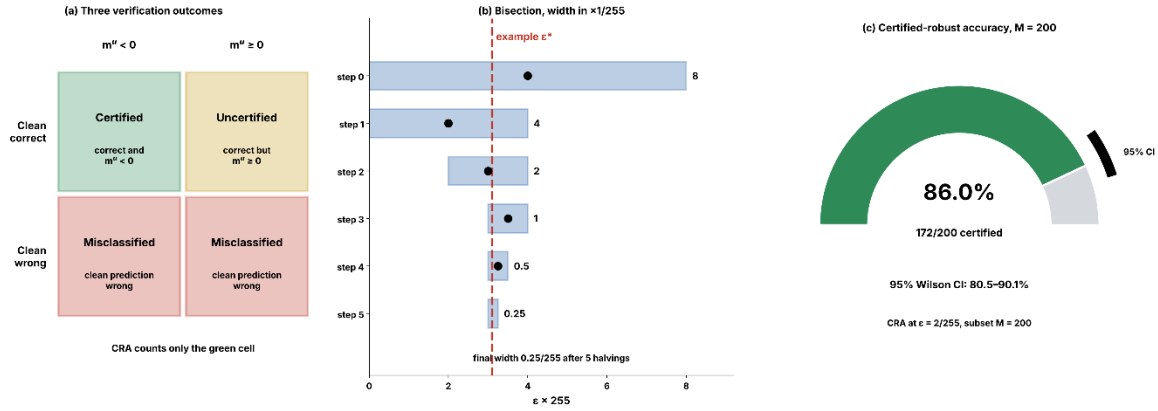

Fig. 8. Verification outcomes, bisection radius search, and certified accuracy.

Fig. 8(a) shows a distinction between certified, uncertified, and misclassified outcomes, with only perfectly classified predictions with a negative upper margin adding to certified-robust accuracy. In panel (b), bisection is used to narrow down the search interval to 8/255 to 0.25/255 in 5 steps, leading to the claimed radius resolution. Only the subset sampling uncertainty is included, and panel (c) shows a 172 to 192 out of 200 to be accurate 86.0% accuracy, with a 95% Wilson interval of 80.5–90.1%. The values for the panels (c) are deduced from the statement for the size of the subset, and panel (b) is simply an example with an arbitrary value for the radius.

H. Ablation Study

To assess each design feature, one feature of SKAC-CertNet is removed or deleted, and the rest of the training configuration is maintained. Results for the accuracy and certified robustness are reported and compared for variants that include the Kolmogorov–Arnold coupling layer, adaptive soft top-k pair selection, saddle-point optimization, and the cone relaxation.

Table 2. Ablation analysis of individual SKAC-CertNet components on PneumoniaMNIST.

Variant	Accuracy (%)	Sensitivity (%)	Specificity (%)	BAcc (%)	F1 (%)	CRA @ 2/255 (%)
Without KAC layer (standard ReLU layer)	91.82	93.27	90.14	91.71	92.65	76.5
Without soft top-k gating (fixed/random pairs)	92.36	93.85	90.62	92.24	93.12	79.2
Without saddle-point training (post-hoc certification only)	93.14	94.23	91.75	92.99	93.68	72.8
LP relaxation instead of sparse SOCP	94.12	95.19	92.47	93.83	95.02	80.6
Full SKAC-CertNet	94.87	96.15	92.74	94.45	95.91	86.0

The individual contribution of KAC coupling, soft top-k gating, saddle-point training, and sparse second order cone relaxation is investigated in Table 3. The final SKAC-CertNet configuration realizes the maximal target classification with certified-robust accuracy, suggesting the possible contributions of structured nonlinear representation, minimax optimisation and convex certification.

G. Comparative Performance Evaluation

This section compares the performance of the proposed SKAC-CertNet with the other approaches for adversarial robustness with respect to certain classification metrics. Accuracy, sensitivity, specificity, and balanced accuracy (BAcc) are all considered when assessing the overall and class-wise reliability of the predictive model. The results confirm the effectiveness of SKAC-CertNet to classify Normal cases and Pneumonia cases and the classification ability remains stable. The comparative analysis also shows that it has more advantages than the traditional ResNet-50 based methods that use LRP and Grad-CAM methods.

Table 3. Comparison of SKAC-CertNet with existing adversarial robustness approaches.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	BAcc (%)
ResNet-50 + LRP (Colin & Surantha, 2025)	91.00	90.00	92.00	91.00
ResNet-50 + Grad-CAM (Colin & Surantha, 2025)	90.60	86.00	83.00	84.50
SKAC-CertNet	94.87	96.15	92.74	94.45

Table 3 shows a comparison between SKAC-CertNet and ResNet-50 based LRP and Grad-CAM approaches on four performance measures. The proposed SKAC-CertNet outperforms the compared approaches in terms of classification effectiveness and class-wise predictive reliability, as it shows the highest accuracy (94.87%), sensitivity (96.15%), specificity (92.74%) and balanced accuracy (94.45%).

H. Discussion

The study filled three gaps in the relatively unexplored area of cone-based verification. The first contribution was replacing post-training certification with objective-based saddle-point training, in which cone relaxation acts as the inner maximizer, allowing model weights and certificates to co-evolve. This design achieved a certified-robust accuracy (CRA) of 86.0%, whereas the design without saddle-point training achieved 72.8%. Second, soft top-k gates were introduced to replace handcrafted coupling heuristics in the Kolmogorov–Arnold coupling (KAC) layer. The contribution of each pair to the dual bound determines gate training rather than weight magnitude. Without soft top-k gating, certified-robust accuracy dropped to 79.2%, while the linear-programming (LP) relaxation variant achieved 80.6%, indicating that the selected cone couplings provide tighter bounds than those obtained through the LP relaxation. Third, conventional MNIST and CIFAR-10 benchmarking was replaced with deterministic certification on PneumoniaMNIST, where small perturbations can potentially affect diagnostic predictions. Certified images are guaranteed to remain correctly classified within the specified perturbation radius under the evaluated threat model. Clean performance was also strong, achieving 95.91% balanced accuracy, 96.15% sensitivity, and 92.74% specificity.

V. Conclusion and Future Work

This research introduced a saddle-point certified training framework for pneumonia detection in chest X-ray images, called SKAC-CertNet. The key innovation is a single minimax pipeline where a sparse second order cone relaxation is used as a differentiable inner maximizer, a Kolmogorov-Arnold coupling layer is used to learn which pairs of neurons should be subject to a cone constraint, and finally, a dual-based margin bound is added directly to the loss. The method is supported by a validity and monotone-tightening lemma and by a convergence analysis of the primal-dual iteration and the exact sparse cone program provides a certificate for each test image. It guarantees that against all attacks with a perturbation size within the clipped $\ell\infty$ ball, an attack is classified correctly. The balanced accuracy was 95.91%, and the certified-robust accuracy was 86.0% at $\varepsilon = 2/255$ on PneumoniaMNIST+ for the full model; the ablation experiments evaluated the contributions of saddle-point training, the KAC coupling layer, soft top-k gating, and cone relaxation. The work hence brings certified robustness closer to the learning process, and provides a fixed benchmark for lightweight biomedical images. These components address the research gap of post hoc verifiers by transforming the certificate into a learning signal and not just an afterthought. Several limits remain. The images are small, the backbone is compact, threat model only includes $\ell\infty$ perturbations, and certification was carried out with a fixed set of test images.

The future work should be extended to the 64×64 size of MedMNIST+, multi-class subset, and 3 dimensions. The verification cost can be reduced with adaptive pair budgets, first-order GPU solvers for the dual problem and other perturbation norms. Deeper backbones will demonstrate the extent of scaling of the approach, and the convergence theory for the non-convex outer problem should be strengthened. The comparison with randomized smoothing and prospective clinical validation will also be planned.

Reference

1. Anderson, B. G., Ma, Z., Li, J., & Sojoudi, S. (2025). Towards optimal branching of linear and semidefinite relaxations for neural network robustness certification. *Journal of Machine Learning Research*, *26*(81), 1–59.
2. Balauca, S., Müller, M. N., Mao, Y., Baader, M., Fischer, M., & Vechev, M. (2024). Gaussian loss smoothing enables certified training with tight convex relaxations. *arXiv Preprint arXiv:2403.07095*.
3. Chiu, H.-M., Chen, H., Zhang, H., & Zhang, R. Y. (2025). Sdp-crown: Efficient bound propagation for neural network verification with tightness of semidefinite programming. *arXiv Preprint arXiv:2506.06665*.
4. Colin, J., & Surantha, N. (2025). Interpretable Deep Learning for Pneumonia Detection Using Chest X-Ray Images. *Information*, *16*(1), 53. https://doi.org/10.3390/info16010053
5. Dhar, J., Goyal, P., Haghighat, M., Zaidi, N., Sohel, F., Vo, Q. B., & Santosh, K. (2026). Towards building robust models for unimodal and multimodal medical imaging data. *Information Fusion*, *127*, 103822. https://doi.org/10.1016/j.inffus.2025.103822
6. Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, & Bingbing Ni. (2024). *[MedMNIST+] 18x Standardized Datasets for 2D and 3D Biomedical Image Classification with Multiple Size Options: 28 (MNIST-Like), 64, 128, and 224* (Version 3.0) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.10519652
7. Kabaha, A., & Cohen, D. D. (2024). Verification of neural networks' global robustness. *Proceedings of the ACM on Programming Languages*, *8*(OOPSLA1), 1010–1039.
8. Ma, L., Chen, J., Qian, L., & Liang, L. (2024). A general approach to improve adversarial robustness of DNNs for medical image segmentation and detection. In O. Colliot & J. Mitra (Eds.), *Medical Imaging 2024: Image Processing* (p. 36). SPIE. https://doi.org/10.1117/12.3006534

9. Maleki, M., Sadig, D., Sidibomma, R., Fazlyab, M., Adibi, A., & Samavi, R. (2026). Certified Robustness via Sparse Second-Order Cone Relaxations. *arXiv Preprint arXiv:2610.16016*.
10. Mao, Y., Balauca, S., & Vechev, M. (2024). Ctbench: A library and benchmark for certified training. *arXiv Preprint arXiv:2406.04848*.
11. Mozhegova, E., Khattak, A. M., Khan, A., Garaev, R., Rasheed, B., & Anwar, M. S. (2025). Assessing the adversarial robustness of multimodal medical AI systems: Insights into vulnerabilities and modality interactions. *Frontiers in Medicine*, *12*, 1606238. https://doi.org/10.3389/fmed.2025.1606238
12. Pal, S., Rahman, S., Beheshti, M., Habib, A., Jadidi, Z., & Karmakar, C. (2024). The Impact of Simultaneous Adversarial Attacks on Robustness of Medical Image Analysis. *IEEE Access*, *12*, 66478–66494. https://doi.org/10.1109/ACCESS.2024.3396566
13. Sarih, A. E., Aneja, N., & Hong, O. W. (2025). Certified Accuracy and Robustness: How different architectures stand up to adversarial attacks. *Intelligent Systems with Applications*, *27*, 200555. https://doi.org/10.1016/j.iswa.2025.200555
14. Shi, Z., Jin, Q., Kolter, Z., Jana, S., Hsieh, C.-J., & Zhang, H. (2025). Neural network verification with branch-and-bound for general nonlinearities. *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 315–335.
15. Singh, A., Sarita, Y. C., Mendis, C., & Singh, G. (2025). Automated Verification of Soundness of DNN Certifiers. *Proceedings of the ACM on Programming Languages*, *9*(OOPSLA1), 1802–1830. https://doi.org/10.1145/3720509
16. Singh, A., Singh, M. P., & Singh, A. K. (2025). Adversarially Enhanced Learning (AEL): Robust lightweight deep learning approach for radiology image classification against adversarial attacks. *Image and Vision Computing*, *154*, 105405. https://doi.org/10.1016/j.imavis.2024.105405
17. Srinivasan, A., Sritharan, D. V., Chadha, S., Fu, D., Hossain, J. O., Breuer, G. A., & Aneja, S. (2026). *Adversarial Robustness of Capsule Networks for Medical Image Classification*. Health Informatics. https://doi.org/10.64898/2026.03.09.26347900
18.
19. Snousi, H. M., & Aleej, F. A. (2025). Machine learning-augmented partial differential equation solvers for high-fidelity engineering design optimization. Journal of Applied Mathematical Models in Engineering, 34-40.
20. Tobler, J., Syeda, H. T., & Murray, T. (2025). A Formally Verified Robustness Certifier for Neural Networks. In R. Piskac & Z. Rakamarić (Eds.), *Computer Aided Verification* (Vol. 15932, pp. 327–348). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98679-6_15
21. Xiao, Y., Ma, S., Zhai, J., Fang, C., Jia, J., & Chen, Z. (2024). Towards General Robustness Verification of MaxPool-Based Convolutional Neural Networks via Tightening Linear Approximation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24766–24775. https://doi.org/10.1109/CVPR52733.2024.02339
22. Zhang, Y., Shen, L., Guo, S., & Ji, S. (2024). Galileo: General linear relaxation framework for tightening robustness certification of transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(19), 21797–21805.
23. Zhang, Z., Wang, B., Yao, L., Keles, E., Jha, D., Antalek, M., Durak, G., Medetalibeyoglu, A., Spampinato, C., Turkbey, B., Gong, B., & Bagci, U. (2026). AdverIN: Monotonic adversarial intensity attack for domain generalization in medical image segmentation. *Medical Image Analysis*, *107*, 103848. https://doi.org/10.1016/j.media.2025.103848
24. Taheri, S., Sabanayagam, M., Ghoshdastidar, D., & Zamani, M. (2026). Robustness Certificates for Neural Networks Against Data Poisoning and Evasion Attacks. *IEEE Open Journal of Control Systems*, *5*, 303–318. https://doi.org/10.1109/OJCSYS.2026.3689712